Claims

1. A process for evaluating an input string to segment said string into component parts comprising:

providing a state transition model based on an existing collection of data records that includes probabilities for segmenting input strings into component parts which adjusts said probabilities to account for erroneous token placement in the input string; and

determining a most probable segmentation of the input string by comparing tokens that make up the input string with a state transition model derived from the collection of data records.

2. The process of claim 1 wherein the state transition model has probabilities for multiple states of said model and a most probable segmentation is determined based on a most probable token emission path through different states of the state transition model from a beginning state to an end state.

3. The process of claim 1 wherein the collection of data records is stored in a database relation and an order of attributes for the database relatioin as the most probable segmentation is determined .

4. The process of claim 3 wherein the input string is segmented into sub-components which correspond to attributes of the database relation.

5. The process of claim 4 wherein the tokens are substrings of said input string.

6. The process of claim 5 wherein the input string is to be segmented into database attributes and wherein each attribute has a state transition model based on the contents of the database relation.

7. The process of claim 6 wherein the state transition model has multiple states for a beginning, middle and trailing position within an input string.

8. The process of claim 6 wherein the state transition model has probabilities for the states and a most probable segmentation is determined based on a most probable token emission path through different states of the state transition model from a beginning state to an end state.

9. The process of claim 5 wherein input attribute order for records to be segmented is known in advance of segmentation of an input string.

10. The process of claim 5 wherein an attribute order is learned from a batch of records that are inserted into the table.

11. The process of claim 6 wherein the state transition model has at least some states corresponding to base tokens occurring in the reference relation.

12. The process of claim 6 wherein the state transition model has class states corresponding to token patterns within said reference relation.

13. The process of claim 8 wherein the state transition model includes of states that account for missing, misordered and inserted tokens within an attribute.

14. The process of claim 13 wherein the state transition model has a beginning, a middle and a trailing state topology and the process of accounting for misordered and inserted tokens is performed by copying states from one of said beginning, middle or trailing states into another of said beginning, middle or trailing states.

15. A machine computer readable medium containing instructions for performing the process of claim 1.

16. A process for segmenting strings into component parts comprising:

providing a reference table of string records that are segmented into multiple substrings corresponding to database attributes;

analyzing the substrings within an attribute to provide a state model that assumes a beginning, a middle and a trailing token topology for said attribute; said topology including a null token for an empty attribute component;

breaking the input record into a sequence of tokens, and

determining a most probable segmentation of the input record by comparing the tokens of the input record with state models derived for attributes from the reference table.

17. A system for processing input strings to segment those records for inclusion into a database comprising:

a) a database management system for storing records organized into relations wherein data records within a relation are organized into a number of attributes;

b) a model building component that builds a number of attribute recognition models based on an existing relation of data records, wherein one or more of said attribute recognition models includes probabilities for segmenting input strings into component parts which adjusts said probabilities to account for erroneous entries within an input string; and

c) a segmenting component that receives an input string and determines a most probable record segmentation by evaluating transition probabilities of states within the attribute recognition models built by the model building component.

18. The system of claim 17 wherein the segmenting component receives a batch of evaluation strings and determines an attribute order of strings in said batch and thereafter assumes the input string has tokens in the same attribute order as the evaluation strings.

19. The system of claim 18 wherein the segmenting component evaluates the tokens in an order in which they are contained in the input string and considers state transitions from multiple attribute recognition models to find a maximum probability for the state of a token to provide a maximum probability for each token in said input string.

20. The system of claim 17 wherein the model building component assigns states for each attribute for a beginning, middle and trailing token position and wherein the model building component relaxes token acceptance by the model by copying states among said beginning, middle and trailing token positions.

21. The system of claim 20 wherein the model building component defines a start and end state for each model and accommodates missing attributes by assigning a probability for a transition from the start to the end state.

22. A string segmentation schema comprising:

a state transition model for a data attribute of a data record wherein the transition model assigns token probabilities to a beginning, middle and trailing state of the model that are transitioned to from a start state and terminate with an end state.

23. The segmentation schema of claim 22 wherein the model copies states amongst the beginning, middle and trailing states to relax token acceptance by said state transition model.

24. The segmentation schema of claim 22 wherein the schema includes a state transition models for multiple attributes of a database record and one or more of said models provide a transition probability between the start state and the end state of each attribute recognition model to accommodate missing attributes within an input string.

25. A process of segmenting a string input record into a sequence of attributes for inclusion into a database table comprising:

considering a first token in a string input record and determining a maximum state probability for said token based on state transition models for multiple data table attributes;

considering in turn subsequent tokens in the string input record and determining maximum state probabilities for said subsequent tokens from a previous token state until all tokens are considered; and

segmenting the string record by assigning the tokens of the string to attribute states of the state transition models corresponding to said maximum state probabilities.

26. The process of claim 25 additionally comprising determining an attribute order for a batch of string input records and using the order to limit the possible state probabilities when evaluating tokens in an input string.

27. A system for evaluating an input string to segment said input string into component parts comprising:

means for providing a state transition model based on an existing collection of data records that includes probabilities for segmenting input strings into component parts which adjusts said probabilities to account for erroneous token placement in the input string; and

means for determining a most probable segmentation of the input string by comparing an order of tokens that make up the input string with a state transition model derived from the collection of data records.

28. The system of claim 27 wherein the state transition model has probabilities for multiple states of said model and a most probable segmentation is determined based on a most probable token emission path through different states of the state transition model from a beginning state to an end state.

29. The system of claim 27 additionally including means for maintaining a collection of records is stored in a database relation.

30. The system of claim 29 wherein the input record is segmented into subcomponents which correspond to attributes of the database relation.

31. The system of claim 30 wherein the tokens are substrings of said input string.

32. The system of claim 30 wherein the input string is to be segmented into database attributes and wherein each attribute has a state transition model based on the contents of the database relation.

33. The system of claim 32 wherein the state transition model has multiple states for a beginning, middle and trailing position within an input string.

34. The system of claim 32 wherein the state transition model has probabilities for the states and a most probable segmentation is determined based on a most probable state path through different states of the state transition model from a beginning state

to an end state.